

CS 595: Assignment #4

Due on Saturday, October 11, 2014

Dr Nelson 4:20pm

VICTOR NWALA

Contents

Problem 1	3
Problem 2	5
Problem 3	6

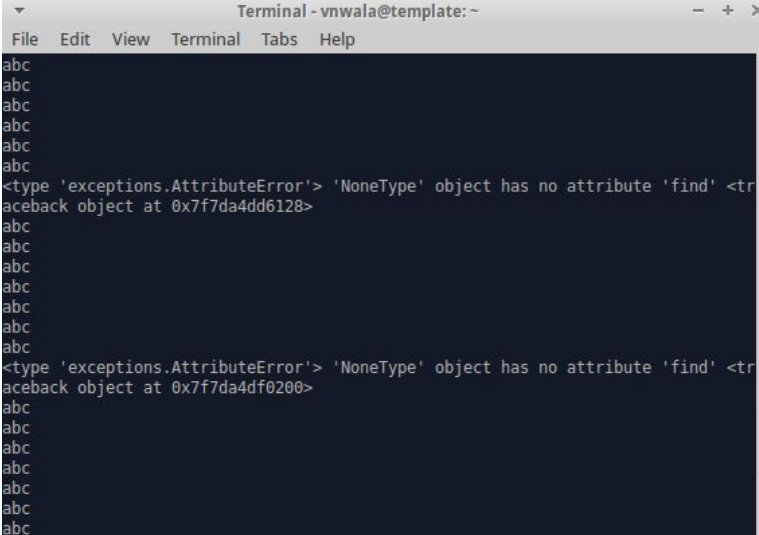
Problem 1

1. From your list of 1000 links, choose 100 and extract all of the links from those 100 pages to other pages. Listing 1 shows a Python script downloading links from input URIs and saving them with their respective md5 encryption names. The files are saved in a folder, URLs.

Listing 1: Script downloading links from input sites.

```
from bs4 import BeautifulSoup
import urllib2
import re
import sys, traceback
5 import hashlib
fh = open("altUrl.txt", 'r')

for line in fh:
    url=line
10    url=url.replace('\n', '')
    try:
        html_page = urllib2.urlopen(url)
        soup = BeautifulSoup(html_page)
        def computeMD5hash(message):
15             m = hashlib.md5()
             m.update(message)
             return m.hexdigest()
        hashMessage = computeMD5hash(url)
        saveFile = open("/home/vnwala/URLs/"+hashMessage+".txt", 'a')
20        saveFile.write('site')
        saveFile.write('\n')
        saveFile.write(url)
        saveFile.write('\n')
        saveFile.write('links')
25        saveFile.write('\n')
        saveFile.close()
        for link in soup.findAll('a'):
            i = link.get('href')
            try:
30                if i.find('http') == 0:
                    print 'abc'
                    saveFile = open("/home/vnwala/URLs/"+hashMessage+".txt", 'a')
                    saveFile.write(str(i))
                    saveFile.write('\n')
35                    saveFile.close()
            except:
                print sys.exc_type, sys.exc_value , sys.exc_traceback
    except:
        print sys.exc_type, sys.exc_value , sys.exc_traceback
```



```
Terminal - vnwala@template: ~  
File Edit View Terminal Tabs Help  
abc  
abc  
abc  
abc  
abc  
abc  
<type 'exceptions.AttributeError': 'NoneType' object has no attribute 'find' <tr  
aceback object at 0x7f7da4dd6128>  
abc  
abc  
abc  
abc  
abc  
abc  
abc  
<type 'exceptions.AttributeError': 'NoneType' object has no attribute 'find' <tr  
aceback object at 0x7f7da4df0200>  
abc  
abc  
abc  
abc  
abc  
abc  
abc
```

Figure 1: extract.py at work

Problem 2

2. Using these 100 files, create a single GraphViz “dot” file of the resulting graph.

I wrote a python script to extract all the URIs within links read from an input file and output the result in a format similar to the dot format, including node labels. I inserted the header and lower brace manually and re-saved the output file in a dot file format.

Listing 2: Script downloading links from input sites saved in a dot format.

```

from bs4 import BeautifulSoup
import urllib2
import re
import sys, traceback
5 import hashlib
from urlparse import urlparse
fh = open("altUrl.txt", 'r')

for line in fh:
10     url=line
    url=url.replace('\n', '')
    try:
        html_page = urllib2.urlopen(url)
        soup = BeautifulSoup(html_page)
15         saveFile = open("gephi_new.txt", 'a')
        saveFile.close()
        for link in soup.findAll('a'):
            i = link.get('href')
            try:
20                 if i.find('http') == 0:

                    saveFile = open("gephi_new.txt", 'a')
                    saveFile.write( '"' + str(i) + "' + '->' + ' ' + url + ' ' + '
                        ;' )
                    k = urlparse(i)
25                     saveFile.write('\n')
                    saveFile.write( '"' + str(i) + "' + '[' + "label" + "=" + k.
                        netloc + "]" + ';' )
                    o = urlparse(url)
                    saveFile.write('\n')
                    saveFile.write( '"' + url + "' + '[' + "label" + "=" + o.netloc
                        + "]" + ';' )
30                     saveFile.write('\n')
                    saveFile.close()
            except:
                print sys.exc_type, sys.exc_value , sys.exc_traceback
        except:
35         print sys.exc_type, sys.exc_value , sys.exc_traceback

```

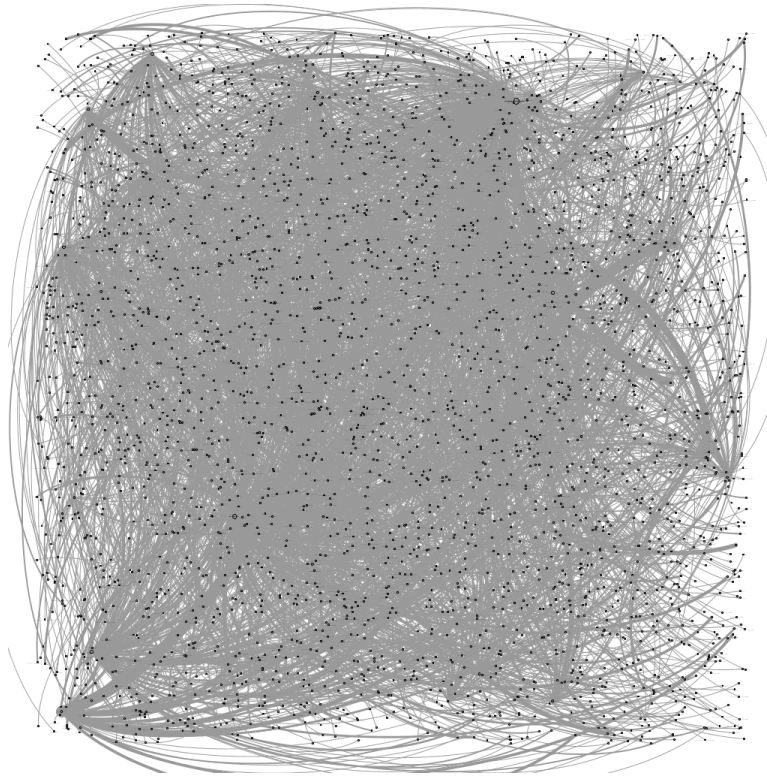


Figure 2: The genaral graph view

Problem 3

3. Download and install Gephi:

<https://gephi.org/>

Load the dot file created in #2 and use Gephi to:

A) visualize the graph (you'll have to turn on labels) B) calculate HITS and PageRank C) avg degree D) network diameter E) connected components

Put the resulting graphs in your report.

For HITS I got two different graphs, the first displayed hubs, the second authority Distribution, with parameter $E=1.0E-4$

The graphs of Betweenness Centrality, Closeness Centrality and Eccentricity Distributions are contain in the parent folder.

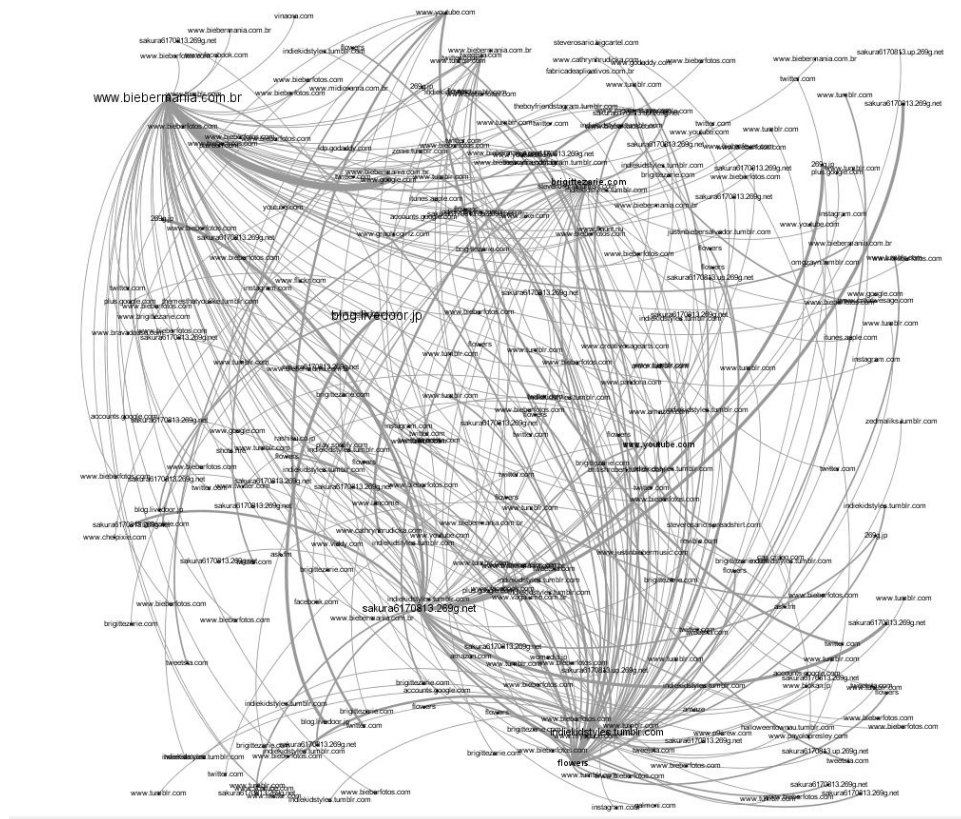


Figure 3: The general graph with 10 percent preview ratio and node labels active.

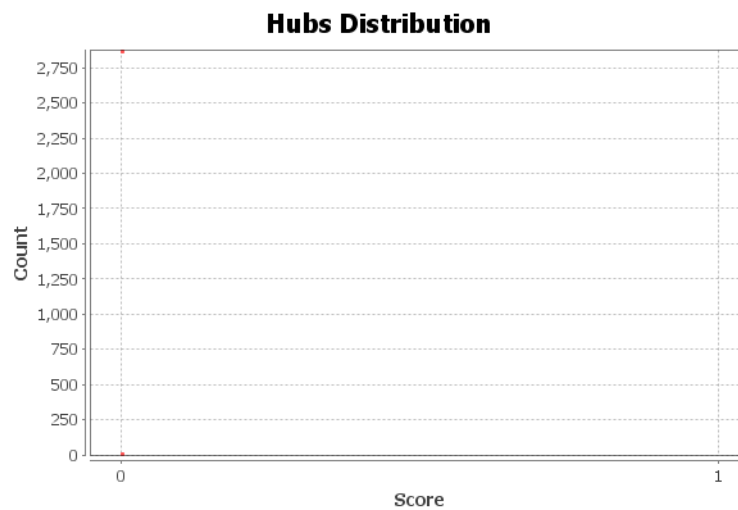


Figure 4: HITS Metric Report

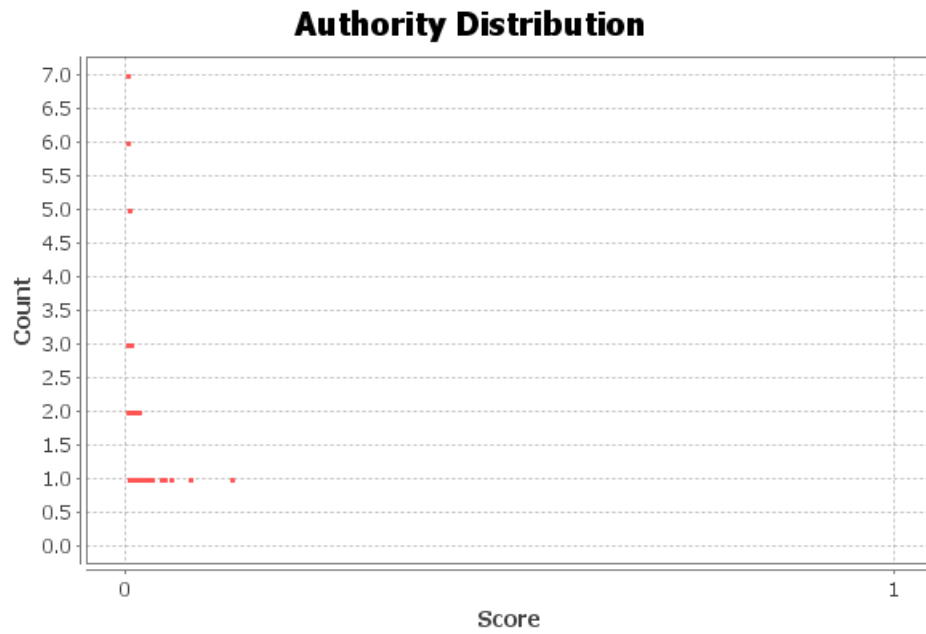


Figure 5: HITS Metric Report

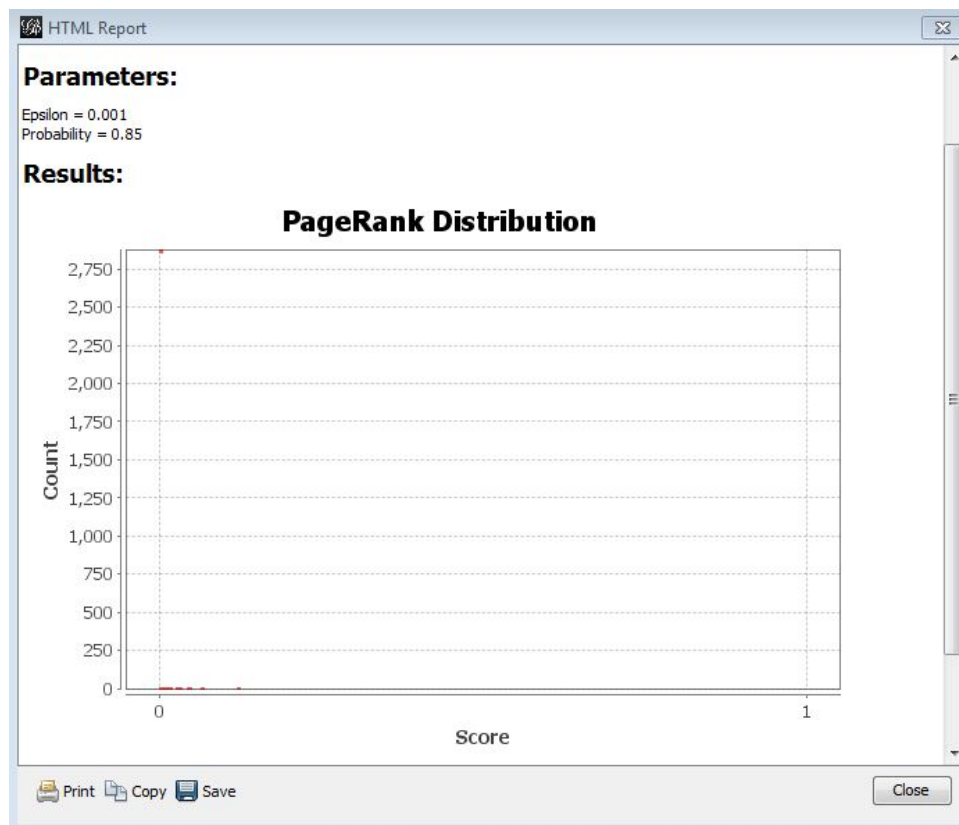


Figure 6: graph of page rank

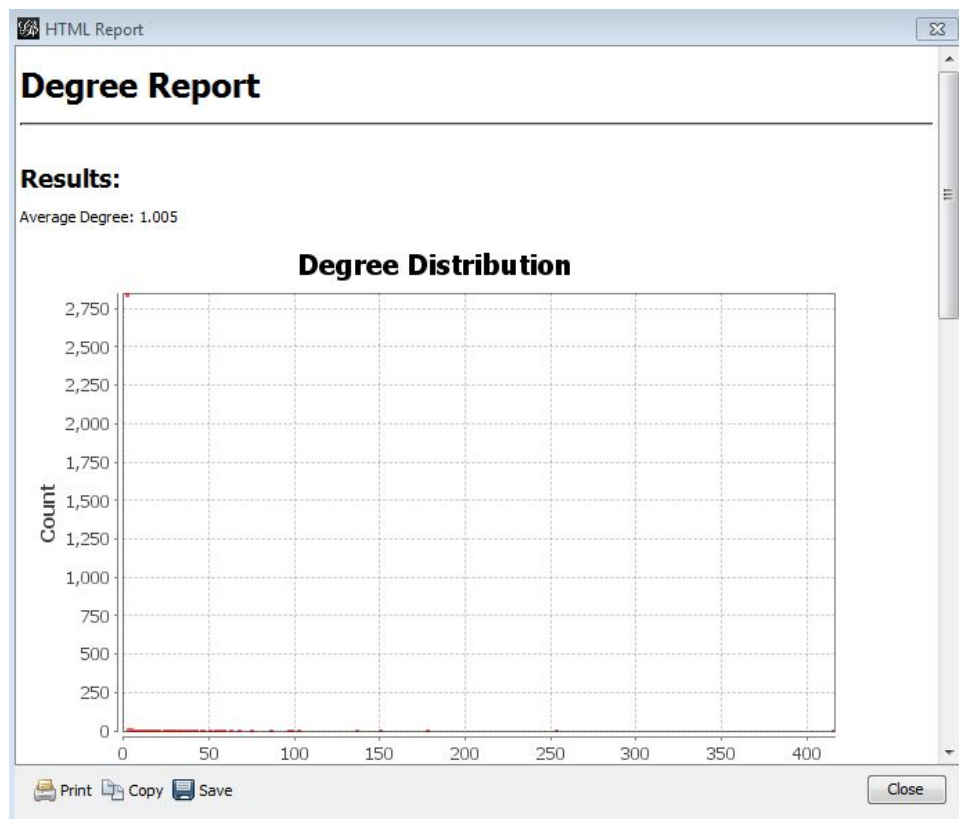


Figure 7: graph of average degree = 1.005

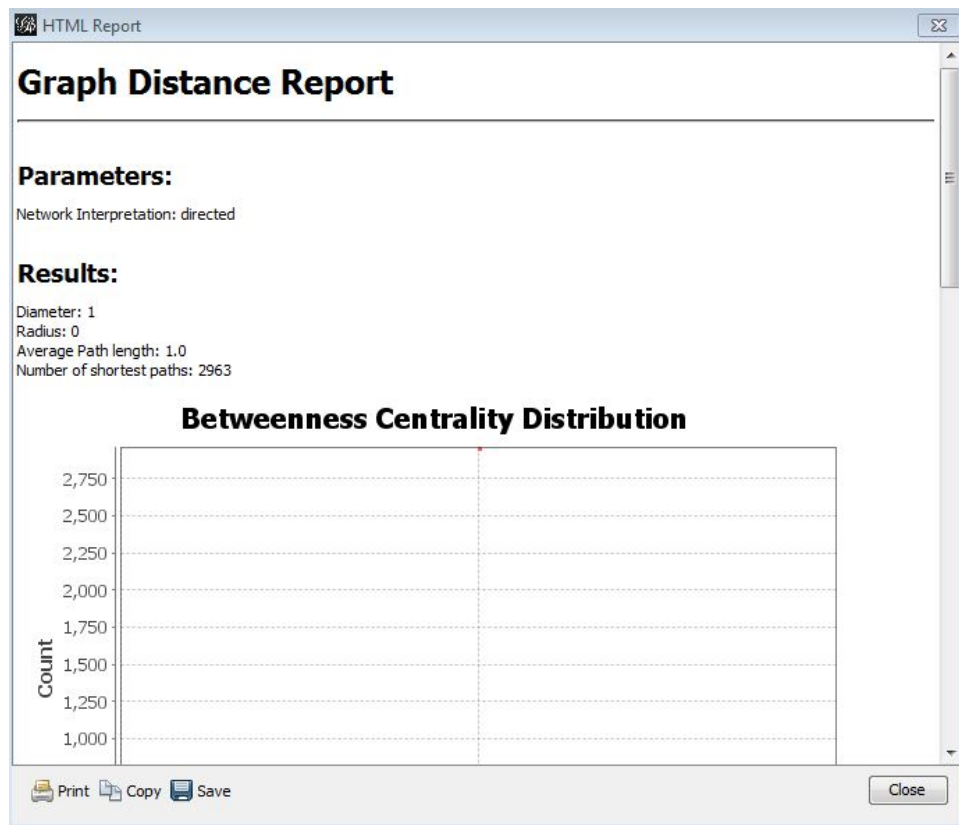


Figure 8: graph of network diameter = 1

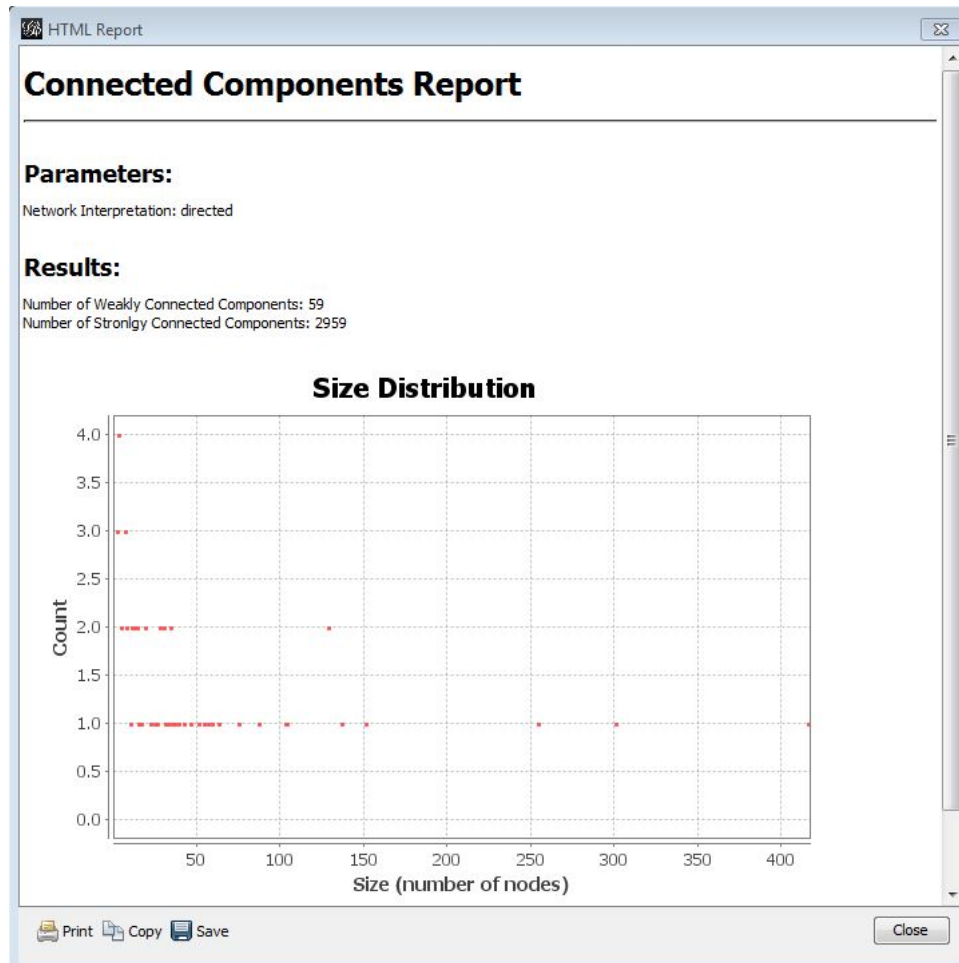


Figure 9: graph of connected components, strongly connected = 2959, weakly connected = 59